> i wanted to understand the contribution of neurons within a specific intermediate layer of llama3.2-1b to the next token prediction task. very interested in seeing how decisions are formed internally.

> utilized the transformer lens library. task was predicting the next token for "she sells sea shells by the", specifically after the final "the" and baseline prediction was "se".

> hooks, a core feature of transformer lens, allow us to inspect and modify a model's internal state during a forward pass. it can read the activation_tensor and modify it. it finds a target neuron and gives an activation value(forced 0 in our case).

> we know the total output of the mlp block (which is then added to the residual stream) is the sum of contributions from all its neurons. and zeroing out in the above step, this sum changes. i/p to the layernnorm preceding the attention mechanism of layer 16 and o/p changes.

> for each neuron in layer 15, i measured if the top predicted token changed from 'sh' and if there's any absolute change in `P('se')` > 0.1 and found 72 neurons out of 8192, showing these changes, which now targets for detailed analysis.

> created a heatmap where brighter cells signify neurons whose ablation causes larger shifts in `P('se')`.
squares marked with 'x' tells that it predicted a completely different token as its top choice.